系統樹最節約復元問題の大域的最適解について(計算理論とその応用)

AUTHOR(S):

成嶋, 弘

CITATION:

成嶋, 弘. 系統樹最節約復元問題の大域的最適解について(計算理論とその応用). 数理解析研究所講究録 1997, 992: 5-11

ISSUE DATE:

1997-05

URL:

http://hdl.handle.net/2433/61163

数理解析研究所講究録
992巻 1997年 5-11

系統樹最節約復元問題の大域的最適解について
– On globally optimal reconstructions of phylogenetic trees –

東海大・理・情報数理　成嶋　弘（Hiroshi Narushima）

最近，著者等によって，植物分類学，動物進化分類学，進化生物学における1960年代以降の成果を背景に，進化系統樹 (以下，単に系統樹と書く) の最節約復元問題に対する数学的に精密な定式化が与えられ，その数理論が展開されている．初めに，これまでのおおよその流れと結果を述べ，つぎに，本論の結果を述べ，最後に，今後の研究の方向と課題を述べる．

1. 最節約問題について

系統樹の最節約復元問題は W. H. Wagner に源を持ち, Farris [1] においてその定式化とアルゴリズムの研究およびコンピュータプログラム開発への第一歩を踏みしめている．系統樹の最節約復元問題とは，いくつかの形質 (表現型) の状態 (値) が観察可能 (計測可能) な化石祖先種および現存種が与えられたとき，系統樹全体の変化量が最小となるように複数の仮想的共通祖先種 (中間種) とそれらの状態および考察対象の種全体の樹形を定めることである．形質 (表現型) の状態 (値) は，伝統的には形態または形態からの計測データによって与えられ，最近は遺伝子解析技術の発展によりDNA からのデータによって与えられることが多い．枝の変化量 (**枝の長さ**と呼ぶ) はその両端点種の形質値差であり，系統樹全体の変化量 (**系統樹の長さ**と呼ぶ) とは各枝の長さの総和のことである．一般には，変化量は形質状態遷移関係と状態間の距離が与えられた形質状態空間のなかで定められる．『系統樹の長さが最小となるように』という基準は 「Wagner Parsimony」または「Maximum Parsimony」と呼ばれている．一般に，複数の形質 (多重形質) について Wagner Parsimony の下での最適な系統樹を求めることになる．

Farris [1] においては，大きく分けて2種類の問題について論じられている．一つは与えられた樹形の下での，形質値の未知な中間種の最適な形質値を求める問題であり，もう一つは最適な樹形と中間種の形質値を同時に与える系統樹を求める問題である．後者の系統樹は Wagner Tree と呼ばれている．我々は前者を**第1最節約復元問題** (The First MPR Problem: MPR は Most-Parsimonious Reconstruction のイニシャルである) と呼び，後者の問題すなわち Wagner Tree を求める問題を**第2最節約復元問題** (The Second MPR Problem) と呼ぶことにする．

2. 第1最節約問題について

Swofford-Maddison [4] は，初めに，Farris [1] において正当性の証明なしで与えられた第1最節約復元問題に対する解法，すなわちいくつかの形質の状態 (値) が既知な化石祖先種および現存種さらにそれら考察対象の種全体を含む樹形が与えられたとき，形質値の未知な中間種の最適な形質値 (形質状態の最節約復元) を求めるアルゴリズム，その不備を補うと共にその正当性の証明を与えている．さらに，その最節約復元は一通りとは限らず複数存在するため，それらのなかで (系統樹解析において) より有用と思われる ACCTRAN 復元と DELTRAN 復元を定めている．ただし，ACCTRAN 復元の源は Farris [1] にある．

Swofford-Maddison [4] においては樹形が “completely bifurcating” (すべての内頂点の次数が 3) の場合が扱われたが，Hanazawa-Narushima-Minaka [8] は樹形が一般の場合の数学的により透明な定式化と共に，Farris-Swofford-Maddison の解法の一般化を与えている．特に，“el-tree (tree with the evaluated leaves)”，“中間 2 点 (median two points)”，“中間区間 (median interval)” 等の概念を導入し，Farris，Swofford，Maddison 等に源をもつ諸概念を論理的に透明なものとすると共に，アルゴリズムやその正当性の証明に必要な諸概念を再帰的に定め，与えられた樹形の下での最節約復元をすべて再帰的に生成する簡明なアルゴリズムを与えている．さらに，各アルゴリズムに対する計算量解析も行っている．その計算量解析は Blum-Floyd-Pratt-Rivest-Tarjan [2] による PICK アルゴリズムに基づいている．

Narushima-Hanazawa [11] は，新たに “中間 4 点 (median four points)” の概念を導入すると共に，各頂点 (形質値の未知の中間種) の MPR set (復元が最節約となるようなその頂点の形質値全体) に関する定理を与えている．特に，その定理を用いることによって，各頂点の MPR set を系統樹の頂点数に関する線形時間で求めることができることを示している．

Narushima-Misheva [12] は，初めに，先の Narushima-Hanazawa method ([8] および [11]) の枠組みの中で，ACCTRAN 復元および MPR-poset (Minaka [7] において最節約復元の間の関係を調べるために導入された最節約復元全体からなる順序集合) を一般の樹形の場合に一般化して透明な定義を与えている．つぎに，Swofford-Maddison [4] に “implicit” に述べられている ACCTRAN 復元の完全最節約性，すなわち ACCTRAN 復元の場合はその系統樹のすべての部分樹も最節約性をもつ唯一つの復元であることを数学的に精密に証明し，さらに ACCTRAN 復元が MPR-poset の最大元となるための特徴づけを行っている．

Narushima [13] は， Narushima-Misheva [12] で証明した ACCTRAN 復元の完全最節約性を ACCTRAN に関する第 1 定理と呼ぶことにすれば，ACCTRAN に関する第 2 定理と呼ぶべき ACCTRAN 復元の形質極値性，すなわち親子の各形質値の増減性とそれらの (各 MPR set の中での) 最大最小性の関係を浮き彫りにする定理を証明し，その定理を用いて， Narushima-Misheva [12] で与えた ACCTRAN 復元が MPR-poset の最大元となるための特徴づけを導いている．

ここで，二三の注意を与えておく．多重形質の場合の系統樹の長さ (その定義は Manhattan distance と呼ばれている) は形質ごとに完全加法的であるから，(樹形が与えられている場合は) 形質ごとに最節約復元を求め，それらを合わせれば多重形質の場合の最節約復元を求めることができる．したがって，第 1 最節約復元問題においては，単一形質の場合を扱えば充分であることがわかる．また，ここまで，形質状態を実数または正の整数として，その線形順序性を暗黙のうちに仮定して述べてきたことを注意しておく．さらに，化石祖先種や現存種など形質状態が既知の種の頂点すべてが求める系統樹の外頂点 (external vertices) となっているという条件の下で定式化し論じてきているが，この条件をはずしてもよい．すなわち，形質状態の既知の種が求める系統樹の内頂点 (intrenal vertices) となっていてもよい．第 1 最節約復元問題においては，後者は容易に前者に帰着することができる．

3. 第 2 最節約問題について

さて，第 2 最節約復元問題すなわち Wagner Tree を求める問題について述べる．初めに，Farris [1] で使われている用語 “tree” と “network” は，それぞれ最近のグラフ理論の用語 “rooted tree” と “tree” に相当することを注意しておく．第 2 最節約復元問題についての Wagner 等の 1960 年代

から現在までの結果は，Hwang-Richards-Winter [6] の Part IV - Chap 2 (Phylogenetic Trees) に Steiner 問題と結びつけ述べられているので，詳しくはそちらを読んでいただくとして，ここでは本論の結果との対比の意味で Foulds-Graham [3] の結果を引用しておく.

For a metric space (S, d), define a weighted graph $G = G(S, d)$ with vertex set S so that each edge $\{s, t\}$ has weight $d(s, t)$. For a finite subset $X \subseteq S$, a *Steiner minimal tree* $S(X)$ for X is a tree having the minimum possible length over all trees in G which contain X in their vertex sets.

It is well known that for arbitrary weighted graphs, finding a Steiner minimal tree (SMT) is in general, an NP-complete problem. It has been shown that for graphs whose edge weights come from certain metric structures, such as the Euclidean plane or the L_1 plane, finding SMTs is also NP-complete.

The problem we are considering, i.e., that of constructing phylogenies, is easily seen to have the following formalization:

For a fixed alphabet A, let d denote the Hamming distance on A^n, that is,

$$d((a_1, \cdots, a_n),\ (b_1, \cdots, b_n)) = \text{ the number of indices } i \text{ such that } a_i \neq b_i.$$

In the metric space (A^n, d), the Steiner problem for phylogeny (SPP) is:

(SPP): Given a set $X \subseteq A^n$, find a Steiner minimal tree $S(X)$ for X.

The following theorem shows that even when A consists of two elements, the SPP for A^n is NP-complete.

Theorem A. *The SPP for $A = \{0, 1\}$ is NP-complete.*

The proof is shown by reducing the known NP-complete problem "Exact 3-Cover" to the SPP (see Foulds-Graham [3]).

Corollary B. *The SPP is NP-complete.*

4. The Main Result

We now describe one theorem and one corollary in this paper. Let V_O be the set of operational taxonomic units (mathematically a nonempty finite set). Let Ω_i be the set of ith character-states (here, the set $\mathbf{R}$ of real numbers or the set $\mathbf{N}$ of nonnegative integers). Let $\sigma : V_O \to \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ (denoted by $\mathbf{\Omega}$) be a function. This function σ is called a *character-state function for V_O*. Let the restriction of σ-range to Ω_i $(1 \leq i \leq n)$ be denoted

by σ_i. Let

$$d(\boldsymbol{a},\ \boldsymbol{b}) = \sum_i |a_i - b_i|.$$

for any elements $\boldsymbol{a}$= $(a_1,\ a_2,\ \cdots;\ a_n)$ and $\boldsymbol{b}$= $(b_1,\ b_2,\ \cdots,\ b_n)$ in $\boldsymbol{\Omega}$. This distance d is said to be *rectilinear*. Then the second MPR problem (2-MPRP) is:

(2-MPRP): Given a set V_O and a character-state function σ for V_O, find an optimal phylogenetic tree T with the set V_O of external vertices evalueted by σ, under Wagner Parsimony criterion.

The optimal phylogenetic tree is called a *globally optimal* solution for the 2-MPRP. The 2-MPRP with σ_i instead of σ is called the *character-wise* 2-MPRP or the 2-MPRP *under a single character.*

Theorem. *The character-wise 2-MPRP can be solved by the computational complexity of sorting the* $|V_O|$ *numbers, and futhermore, the solution is essentially unique.*

We here illustrate the theorem by using an example. Let $V_O = \{f,\ g,\ h,\ i,\ j,\ k,\ l\}$, where f is a unique root such as a species of fossil, and $\{g,\ h,\ i,\ j,\ k,\ l\}$ is a set of present day speices. Let a character-state function σ for V_O be given in 表 1.

表 1: V_O and σ

v	f	g	h	i	j	k	l
$\sigma(v)$	$(1,2,2)$	$(3,0,1)$	$(0,1,1)$	$(6,2,3)$	$(5,3,0)$	$(2,4,2)$	$(4,5,0)$

For example, $d(\sigma(f),\ \sigma(g)) = |1-3| + |2-0| + |2-1| = 5$. A globally optimal solution for the (character-wise) 2-MPRP on the first character is found as follows:

Step 1. Sort in ascending order the following vertices (OTUs) with the known character-states, according to the character-states:

$$(f,1),\ (g,3),\ (h,0),\ (i,6),\ (j,5),\ (k,2),\ (l,4).$$

Then we have the following:

$$(h,0),\ (f,1),\ (k,2),\ (g,3),\ (l,4),\ (j,5),\ (i,6).$$

Step 2. Construct a tree shown in 図 1.

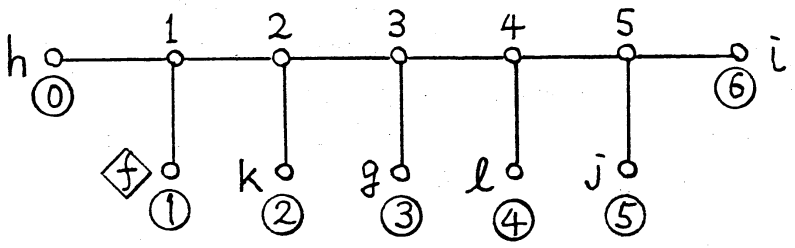

図 1: A tree

Step 3. Root the tree in 図 1 at f. Then we have a globally optimal solution shown in 図 2. The length of the rooted tree is 6.

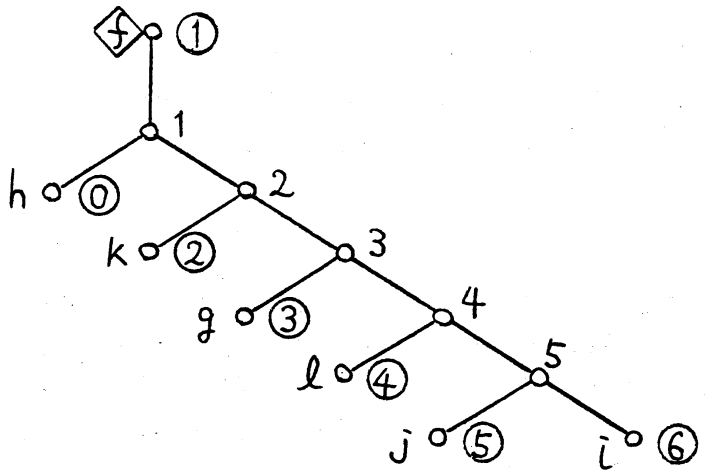

図 2: a globally optimal solution on the first character

A globally optimal solution on the second character is similarly shown in 図 3.

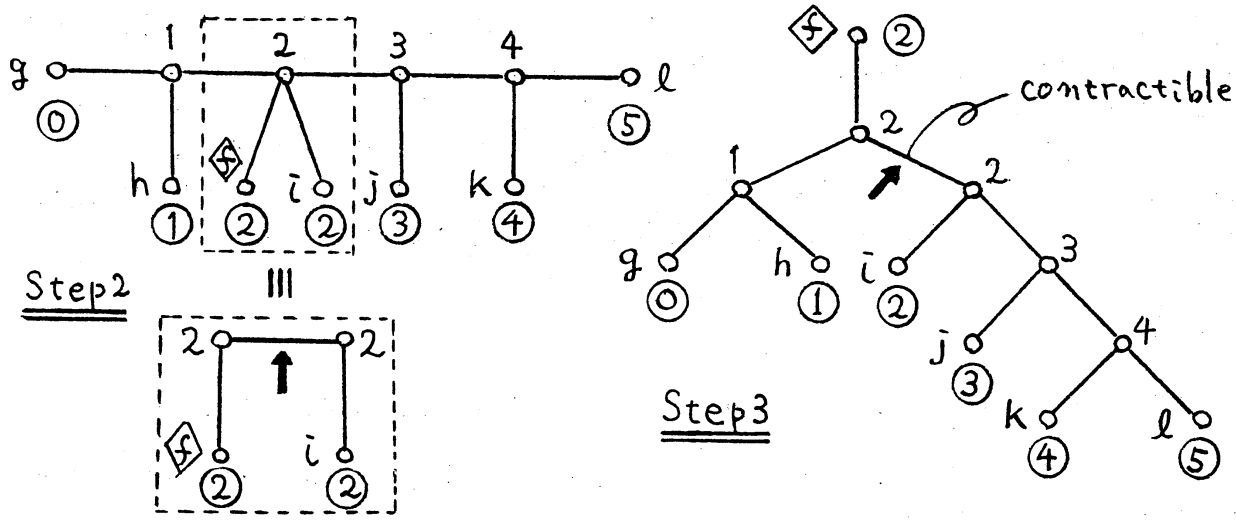

図 3: a globally optimal solution on the second character

The length of the rooted tree is 5. The relation $\equiv$ means that this part may be any tree with the external vertices f and i having the same character-state. Note that other parts may have similar variations. Any two trees equivalent under the relation $\equiv$ are said to be *state-homogeneous.* The "is essentially unique" in the theorem means "is unique up to state-homogeneity". The fact that the tree-topology of the rooted tree in 図 2 and that of the rooted tree in 図 3 are different, is a critical point.

Corollary. *Let $(\Omega, \leq)$ be a poset with the usual ordering. Then if $\sigma(V_O)$ is a chain in $(\Omega, \leq)$, the 2-MPRP can be solved by the computational complexity of sorting the $|V_O| \times p$ numbers for p = the number of characters, and futhermore, the solution is essentially unique.*

Note that comparing the SPP and the 2-MPRP, we have a difference such that some elements in "X" of the SPP may be internal vertices of the solution tree, but all elements in "V_O" must be external vertices of the solution tree.

5. 今後の研究の方向

問題の数学的に精密で透明な定式化により，解法や解法の計算量が明らかになると共に，諸概念の定性的性質やそれらの関連および構造も浮き彫りになり，数理論としてのととのいをみせつつある．進化生物学における『最節約原理』という簡明な一原理の数学的定式化の結果，その世界に潜む数理的内容が明らかになるにつれ，特に，離散数理的性質が豊富で，しかも論理的に簡明な美しい世界に対し感銘を禁じえないものがある．

今後の課題として，最節約復元のなかで ACCTRAN 復元と並んで重要な DELTRAN 復元の特徴付け，MPR-poset 上での ACCTRAN 復元と DELTRAN 復元の関係，MPR-poset の束論的構造，および第 2 最節約復元問題の一般的解法等の多くの研究課題がある．また，形質状態のより一般的な遷移関係の下での研究は，その一例として Swofford-Maddison [5] など種々あるが，この場合の数理的により精密な展開は今後の課題の一つでもある．これらの課題に取り組み，理論としての進化発展をはかると共に，得られたアルゴリズムのコンピュータインプリメントも行い，現在すでにあるソフトウェアとの関連の下に，最節約復元問題のソフトウェアの開発作成も課題の 1 つである．さらに，我々の数理論が実際の問題でどのような意味を持つのか，すなわち，進化生物学上の諸事象の解明にどのように有効であるのかを考察することも重要な課題であろう．

参考文献

[1] J. M. Farris, Methods for computing Wagner trees, Systematic Zoology 19 (1970) 83-92.

[2] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, Time bounds for selection, JCSS 7 (1973) 448-461.

[3] L.R.Foulds and R.L.Graham, The Steiner problem in phylogeny is NP-complete, Advances in Applied Mathematics 3, (1982) 43-49.

[4] D. L. Swofford and W. P. Maddison, Reconstructing ancestral character states under Wagner parsimony, Mathematical Biosciences 87 (1987) 199-229.

[5] D. L. Swofford and W. P. Maddison, Parsimony, character-state reconstructions, and evolutionary inferences [in Systematics, Historical Ecology, and North American Freshwater Fishes (ed. R. L. Mayden), Stanford Univ. Press, 1992].

[6] F.K.Hwang, D.S.Richards and P.Winter, The Steiner Tree Problem (Annals of Discrete Mathematics 53), North-Holland, 1992.

[7] N. Minaka, Parsimony, phylogeny and discrete mathematics: combinatorial problems in phylogenetic systematics (in Japanese: with English summary), Natural History Research, Chiba Prefectural Museum and Institute, Vol.2 No.2 (1993) 83 - 98.

[8] M. Hanazawa, H. Narushima and N. Minaka, Generating most parsimonious reconstructions on a tree: a generalization of the Farris-Swofford-Maddison method, Discrete Applied Mathematics 56 (1995) 245-265.

[9] 成嶋 弘, 進化生物学における離散最適化問題の解法について– 祖先形質復元問題に対する線形時間アルゴリズム –, 数解研講究録 950『計算モデルと計算の複雑さに関する研究』(1996 年 5 月) 46-55.

[10] H. Narushima and N. Misheva, On a role of the MPR-poset of most-parsimonious reconstructions in phylogenetic analysis – A combinatorial optimization problem in phylogeny –, Proc. The International Symposium on Combinatoircs and Applications (eds:W.Y.C.Chen, D.Z.Du, D.F.Hsu, H.Y.Hap) (June 28 - 30, 1996 : Nankai Institute of Mathematics, Nankai University, Tianjin, P.R.China) 306-313.

[11] H. Narushima and M. Hanazawa, A more efficient algorithm for MPR problems in phylogeny, (to appear).

[12] H. Narushima and N. Misheva, On the characteristics of the ancestral character-state reconstruction under the accelerated transformation optimization, to appear.

[13] H. Narushima, On most-parsimonious reconstruction in phylogeny and extremal properties of ACCTRAN reconstructions, to appear.